CAS ETH in Machine Learning in Finance and Insurance

Yield Crop Estimation Based on Remote Sensing Data

Project report

Tomasz Skorkowski

05th November 2024

Abstract

We investigate the problem of crop yield prediction using remote sensing data. We look at the problem from the perspective of insurance company developing an insurance product to support farming communities. Consequently we are not only interested in low average prediction error but model performance when crop yield is low as well.

We follow the approach proposed by Jiaxuan You et al. in "Deep gaussian process for crop yield prediction based on remote sensing data." [1]. We apply three deep learning architectures: multilayer perceptron, Long Short Term Memory and Convolution Neural Network and compare it with ridge regression. For training we use histograms derived from Sentinel2 top of the atmosphere reflectance satellite images. We train and evaluate our models on the county-level corn crop yield in the U.S. in the years 2016-2022.

Introduction

Problem description

Predicting crop yield is an important task from multiple perspectives. On the global and administrative level it is relevant for food security. Under the impact of growing population and climate change it is expected that globally, food demand will increase by 35% to 56% between 2010 and 2050. At the same time it is estimated that the population at risk of hunger will either decrease by 91% or increase by 30% [6]. On the local level accurate estimation of crop yield helps to select the best season and plants to grow.

There are multiple risk factors that affect the crop yield and that are difficult to manage on the individual level and hence create an opportunity for insurance and reinsurance companies to support and help distribute risk on the national and global level.

The primary aim of this project is to apply machine learning models to satellite images to predict corn crop yield in the USA. The secondary objective is to support the strategic goal of Blue Marble to build knowledge and predictive models that would allow extension of insurance coverage beyond weather risks. As parametric insurance is the main business driver of this investigation, model performance for poor harvest years is of special interest.

Blue Marble is an Impact InsurTech with a mission to bring insurance to the underserved, farming communities. They are present in Latin America, South and South-East Asia and Africa.

Literature discussion

Due to its importance, crop yield prediction is a popular research topic. In recent years various machine learning methods have been applied to the problem. Systematic literature reviews from 2022 [4] and 2020 [5] showed a growing interest in AI driven methods. According to both studies the most popular architectures for the task are CNN and LSTM.

In the article [1] authors propose the use of histograms of pixel counts as input features into LSTM and CNN models to predict soybean crop yield in the U.S.. Due to the popularity of those two architectures and the attractive dimensionality reduction property of histograms it is an attractive proposition as a starting point for future research. For exactly this reason the decision has been made to attempt to replicate machine learning and data models proposed by the authors.

Literature review for this project revealed two additional papers that could be particularly interesting as a future extension due to interesting application of transformers.

- In [3] the author proposes a new attention mechanism designed to work with multispectral satellite images. In the paper satellite images are used as features for land cover classification
- In [2] authors develop a multimodal transformer model utilising both satellite images and weather data for crop yield prediction.

Modeling

High level description

Following [1], we make an assumption that the location of a pixel in an image is not as important as its color. This allows us to reduce dimensions of the data significantly, group pixel values together and work with the counts of pixels i.e. histograms rather than full images themselves.

To make data collection possible within the timeframe of the project further simplifying assumptions is that enough information will be captured at pixel resolution of 60x60 meters and that it is possible to reduce corn season lasting between May and September to three two-month periods that can be represented by the median value of each pixel. It means that there are:

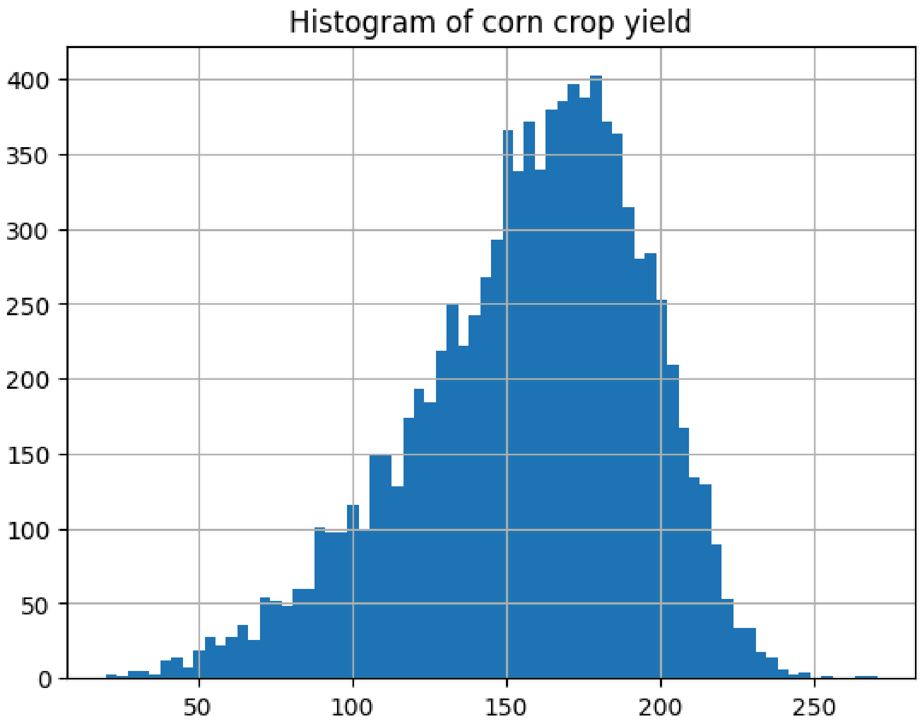
- 3 images per year,
- Each pixel has a resolution of 60 by 60 meters
- Pixel's value is a median value of all images that had less than 35% of cloudy pixels captured by Sentinel2 satellite within each two month period.

Data collection

Labels were obtained from CropNet[1] dataset that was recently published on Hugging Face. The source of the labels is the United States Department of Agriculture and contains information on yearly corn yield measured in bushels per acre for the years 2016-2022. We choose the finest level of aggregation possible and target predictions on the county level.

The distribution of labels is slightly skewed to the left with mean of 157, standard deviation of 38 and skewness of -0.58. Given that skewness is not significant we have decided not to apply any transformation to the data.

[1] CropNet/CropNet · Datasets at Hugging Face



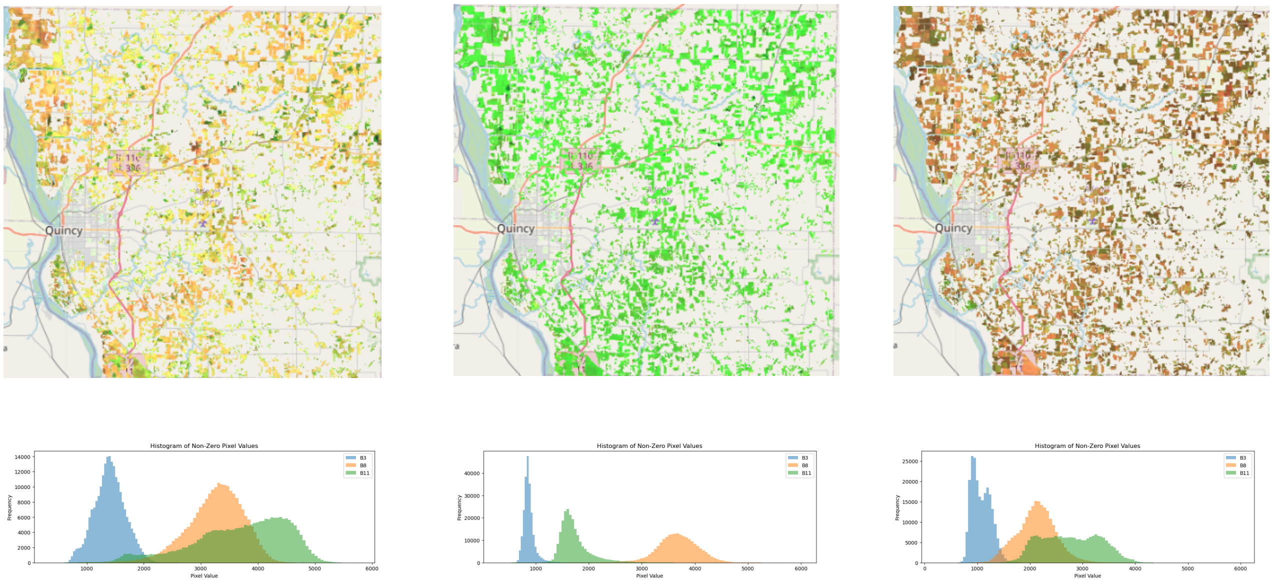
We have collected 28,661 Sentinel2 satellite images via google earth engine. To obtain a dataset that is in line with the labels, each image is cropped according to the county administrative boundaries based on 2018 census data and pixels representing corn crop according to the USDA National Agricultural Statistics Service. After cropping the images take roughly 80gb of storage space.

We transform images by separating each image channel into 60 intervals of equal length. We count pixel values falling into each interval and construct a histogram. One datapoint is a concatenation of 3 histograms, together they capture the information from the whole harvest season. In case one of the three histograms for the specific county and year combination were missing we used an empty histogram as a placeholder. Conversely, when two or more histograms were missing, the sample was discarded.

In total 9,578 data points have been prepared for regression. Histograms used for training are based on 9 channels: blue, green, red, red edge 1, red edge 2, red edge 3 and red edge 4 as well as near infrared and water vapor. Each histogram has 60 bins. These parameters were chosen after experimenting with different values. Lower levels lead to poor model performance and going beyond did not provide much improvement while in extreme cases also lead to worse model performance.

Sample satellite images

Below, a sample of three satellite images overlaid with a map of the USA. The images show Adams county in Illinois in 2017. From the left, median of months May and June, July-August and September-October and their corresponding histograms. Images and histograms are visualisations of three channels: green, short wave infrared and near infrared.



2

Given that different seasons are clearly distinguishable the initial assumptions about importance of pixel values seem plausible. Perhaps even more so when we consider that each image has 13 channels in total.

Models, training and evaluation

We set aside 20% of data to use as a test dataset for model evaluation. We split the remainder further into training and validation datasets. Each holding respectively 64% and 16% of the initial data.

Four separate model architectures were implemented for comparison. Simple, fully connected multi layer perceptron, LSTM, CNN and ridge regression as a baseline.

[2]

General observations

Based on loss function only there is no clear favorite among the three approaches.

MLP has proven to be very flexible and can easily work with a penalized loss function. We tried two weighted versions of mean-squared error. One applying a higher penalty below the low yield threshold of 80 and a lower penalty above the high yield threshold. The other applied a progressively higher penalty for yield below the low threshold only. Slightly better results for predicting bad years comes with higher variance on the unseen data.

LSTM architecture has proven to be difficult to work with. Most changes in the model architecture beyond adding an attention layer leads to the vanishing gradients and constant predictions. Having said that, once the model finds the good path beyond the average the model offers the smoothest learning path.

CNN had the advantage of being quite flexible without convergence issues. It has decent tail performance and lower variance in predictions.

Interestingly, introducing a covariate representing information about the state that the county belongs to improved performance of LSTM and MLP. Due to differences in architecture and shape of the inputs the same was not tested on CNN. Before the introduction of state code CNN model was performing marginally better than the other two. With the extra covariate however LSTM generalizes better and performs better in low yield scenarios. The improved model performance with additional covariate indicates that there are other important factors, potentially linked to the geography that are not covered by satellite images.

For brevity only LSTM architecture will be presented in detail. All experiments performed during this project are available for inspection at weights and biases[3] portal. Furthermore github repository[4] contains the codebase necessary to replicate the experiments. Access to both is public.

LSTM model architecture

Model layers:

- 7 stacked LSTM layers with 200 units each

[3] https://wandb.ai/t-skorkowski/blue-marble?nw=nwusertskorkowski

[4] https://github.com/tskorkowski/crop_yield_prediction_CAS

- Recurrent activation function: sigmoid
- Activation function: hyperbolic tangent
- Additive attention layer with 200 attention units
- Fully connected linear output layer

This LSTM model implementation has approximately 8 million parameters.

Training parameters:

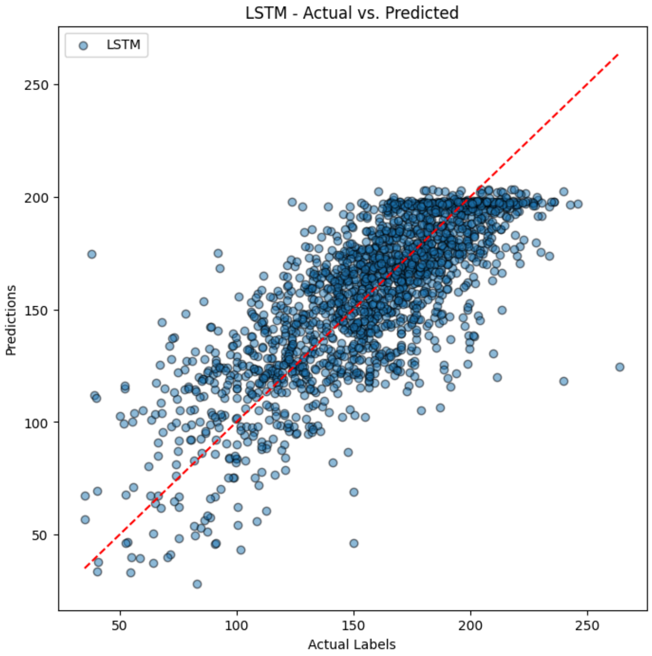
- Learning rate: 0.00013
- 300 epochs with callback monitoring validation loss

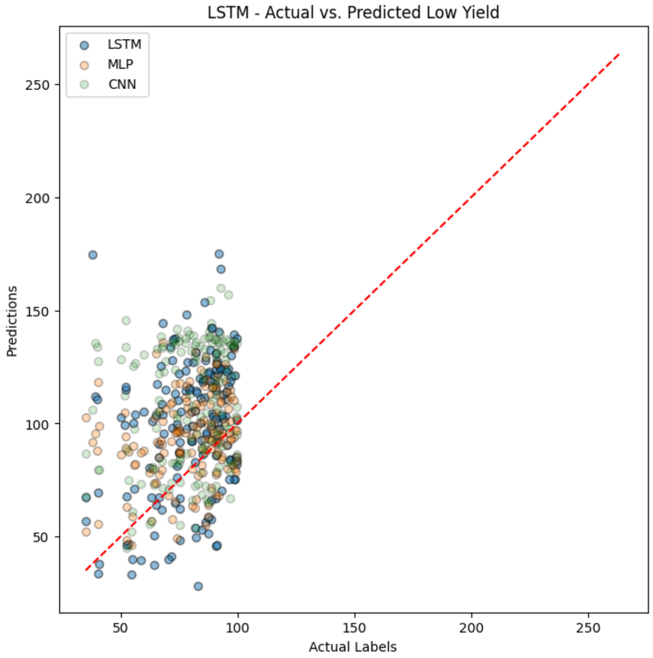
Model performance on test data

Model	RMSE	MAE	Corr
LSTM	24	19	0.777
CNN	26.6	19	0.759
MLP	32.64	22	0.75
Ridge Regression	39.48	26.89	0.513

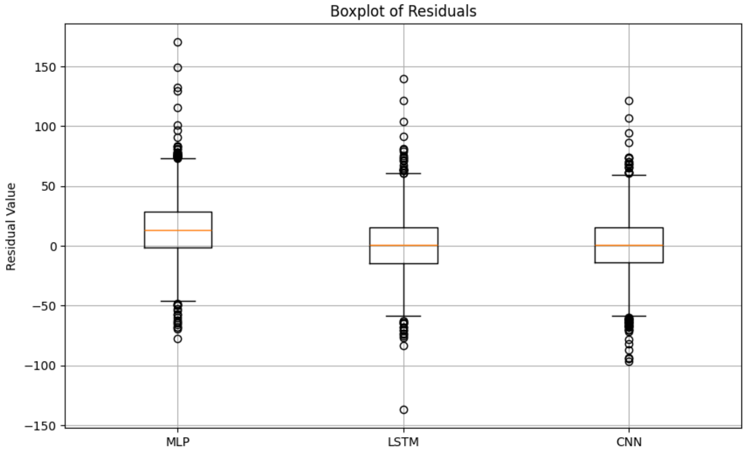
If we take 100 bushels per acre as a low crop threshold then 8.4% or 158 samples fall into that category. Mean squared error for observations below this threshold is the lowest for MLP and equal to 7.45, followed by LSTM and 10.4 and finishing with CNN and RMSE of 12.

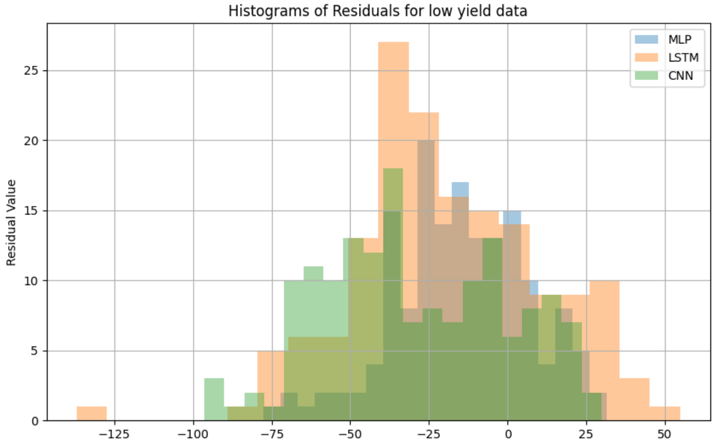
Looking at residuals and model fit we observe that even though residuals are centered around the zero, all models tend to significantly overestimate low crop yields and underestimate high crop yields. From a practical perspective that means that the models are not good enough yet to build an insurance product. This however could be rectified by relaxing assumptions, collecting more images per year and also including more recent years in the data.





With current models benefit payments would only sometimes be made to those who need it. MLP, the best performing model in the left tail correctly predicts low yield threshold only in 50% of cases.





Since authors of reference study work with soybean crops instead of corn we lack direct ability to compare results. Nevertheless, qualitatively it seems that our models perform significantly worse. Among other things, this could be due to difference in channels captured by MODIS satellites or due to difference in temporal data - data analysed in the paper ends prior to 2017 where our dataset starts. Similarly the authors were interested mostly in average prediction error rather than specific behavior in lower quantiles. Having said that, the obtained results further confirm the value of satellite images as a data source and serve as a starting point for future investigations.

Next steps

The results obtained so far seem to indicate that there are multiple options on how to take this work forward. One way would be to capture higher resolution data. Due to time constraints this was impossible for this project but resolution can easily be increased to 10x10m or 20x20m meters, depending on the channel. Furthermore, the impact of capturing more satellite images per year and data augmentation techniques could be studied in the future.

Another natural extension to the current work would be to apply CNN directly to the images instead of histograms. Also, reduction of a research question to classification problem could be more beneficial from the perspective of building a simple insurance product.

Yet another attractive idea would be to replicate the two prospective papers mentioned in the literature discussion. That approach not only uses more recent model architecture but takes advantage of multiple data modalities as well.

Acknowledgements

I would like to thank Blue Marble’s team for suggesting this inspiring topic, support and conversations. In particular Marco Bacci, Jon Otegi, Claudio Valletta and Julian Widmer.

References

[1] You, J., et al. “Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data.” Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-17), vol. 31, no. 1, 2017.

[2] Lin, F., et al. “MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer.” Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[3] Rad, R. “Vision Transformer for Multispectral Satellite Imagery: Advancing Landcover Classification.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.

[4] Muruganantham, P., et al. “A systematic literature review on crop yield prediction with deep learning and remote sensing.” Remote Sensing, 2022.

[5] Van Klompenburg, T., et al. “Crop yield prediction using machine learning: A systematic literature review.” Computers and electronics in agriculture, vol. 177, 2020.

[6] Van Dijk, M., et al. “A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050.” *Nature Food*, 2021.